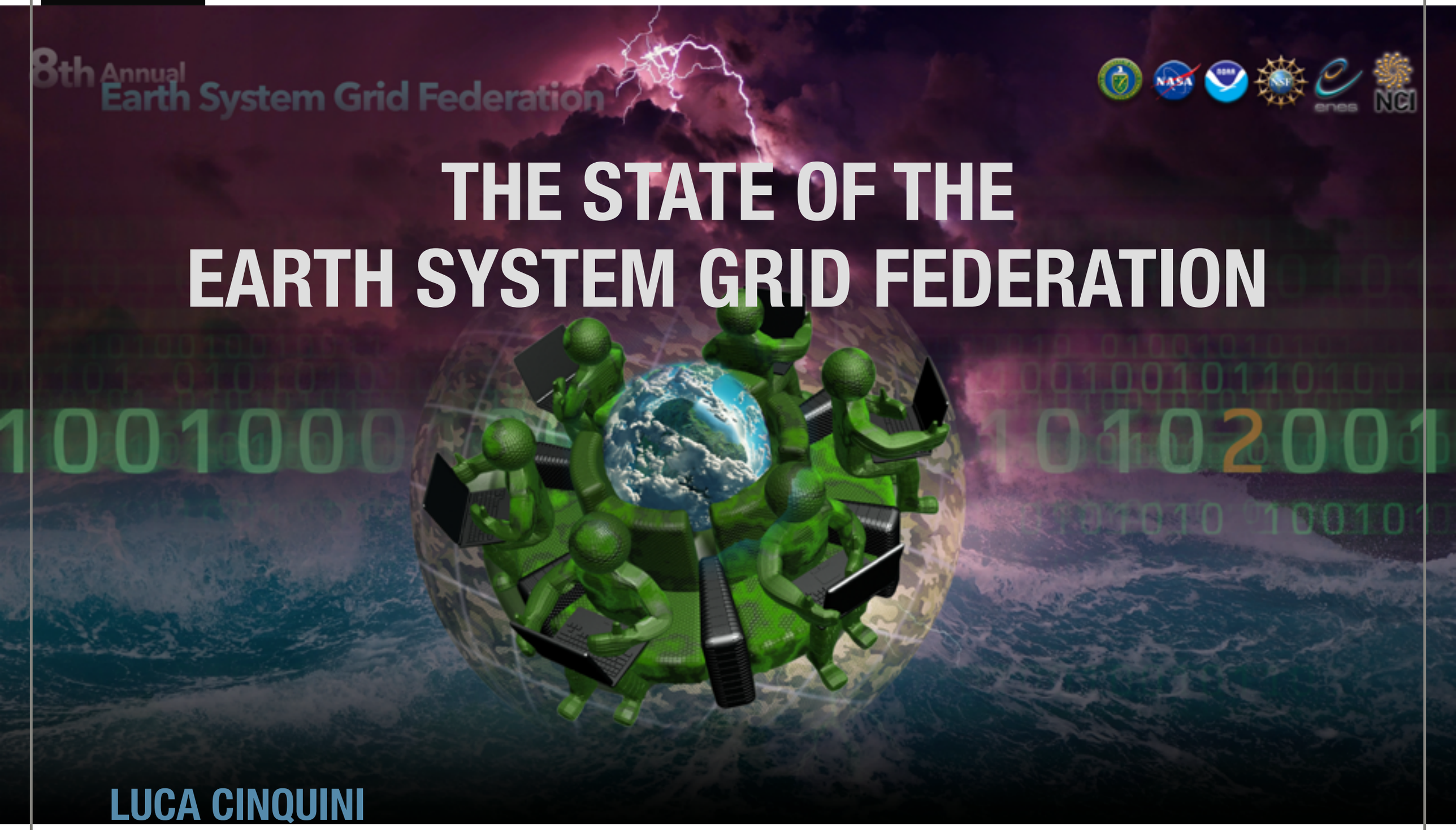


THE STATE OF THE EARTH SYSTEM GRID FEDERATION

LUCA CINQUINI

NASA JET PROPULSION LABORATORY AND CALIFORNIA INSTITUTE OF TECHNOLOGY

JPL UNLIMITED RELEASE SYSTEM CLEARANCE NUMBER: #

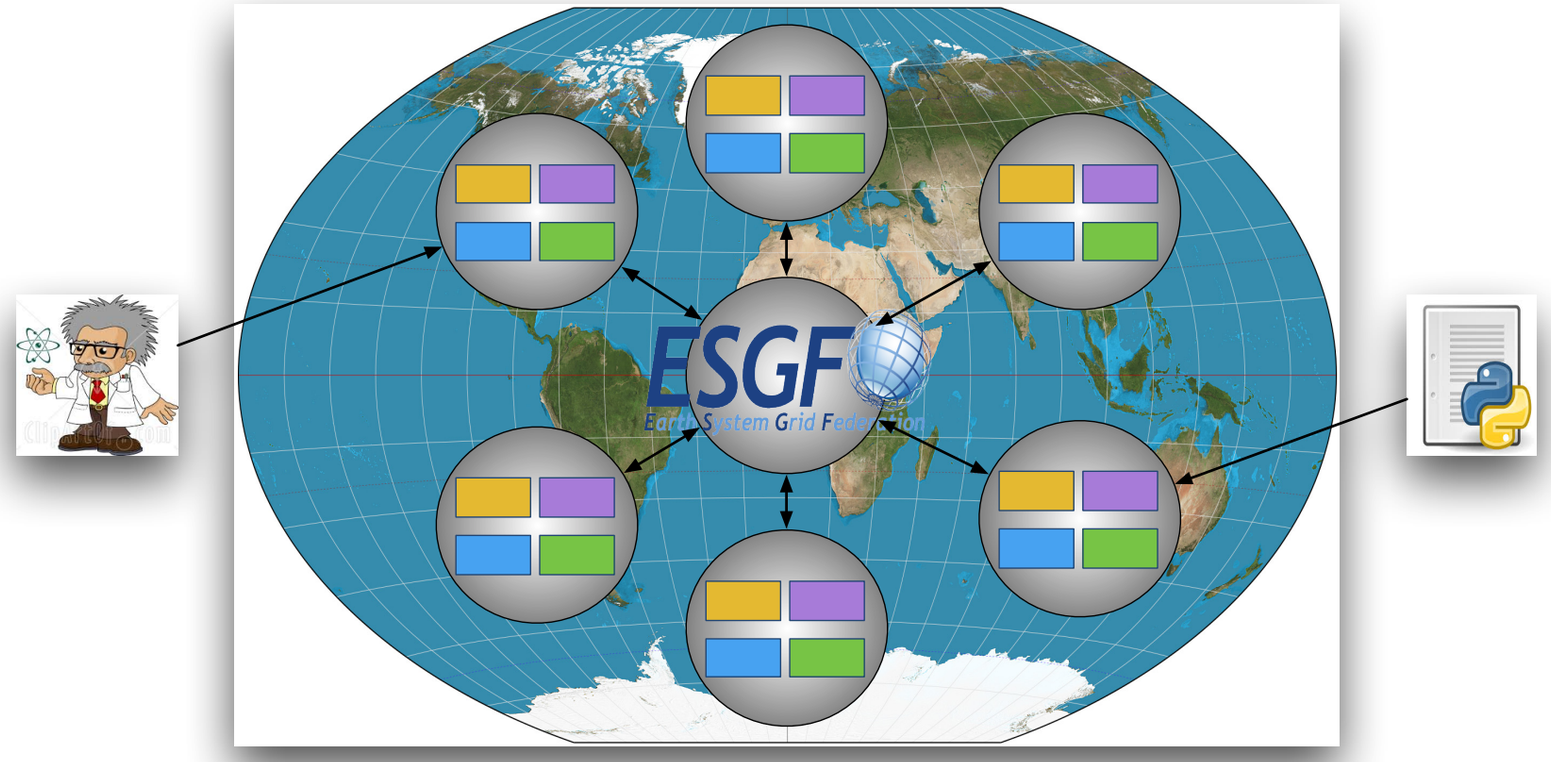
Introduction

- Climate change is one of the most serious threats presently facing the Earth ecosystem:
  - New report indicates climate change is accelerating beyond earlier predictions
  - We are the last generation that can avoid a massive extinction of species
- As the world leading data infrastructure in support off climate change research, ESGF plays an important role in helping to predict and alleviate drastic climate changes

grizzly + polar = “grolar” or “pizzly”

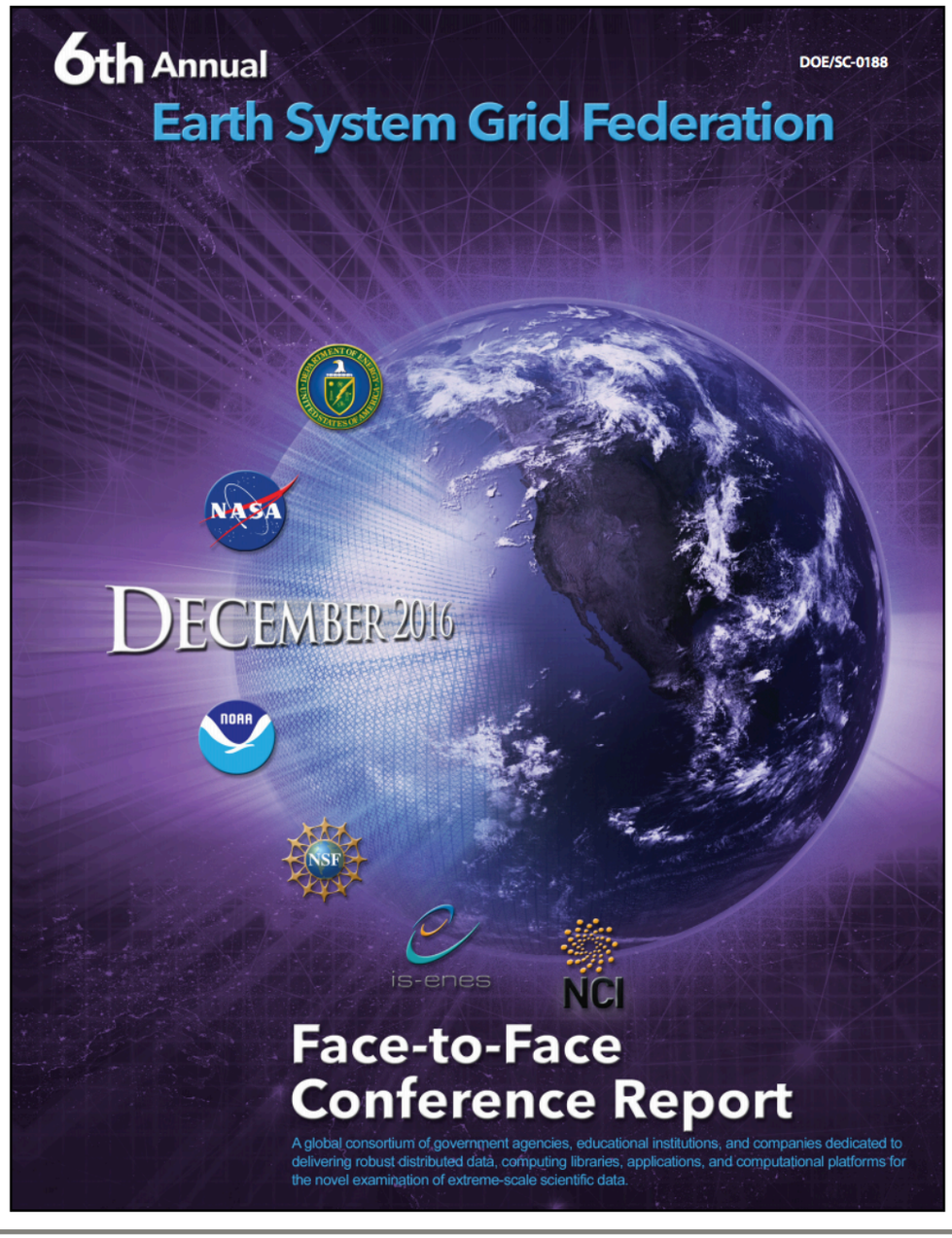
State of ESGF

- ESGF has made constant, solid progress in 2018: improving the reliability of the technical infrastructure, developing new functionality, expanding data holdings and user base
- Most recent stats: 31 nodes, 793,026 datasets, 10,054,190 files, 133 CoG projects, 19,978 users
- …but no “R&D 100” Award in 2018…



Review of Action Items from 2017 F2F

- Roadmap established by ESGF-SC and ESGF-XC in 2017 for 2018 (https://esgf.llnl.gov/esgf-media/pdf/2017-ESGF_F2F_Conference_Report.pdf):
  - Short Term Plans for "CMIP6 Preparedness" (0-2 years):
    - Replication
    - Documentation and training for data publishers
    - Software and operations security
    - PID Service
    - Basic data reduction and analysis operations
    - User authentication and authorization
  - Longer Term Plans for ESGF longevity (2-5 years):
    - Server-side computation
    - Installation
    - Cloud computing
    - Programmatic access to data



2018 ACCOMPLISHMENTS

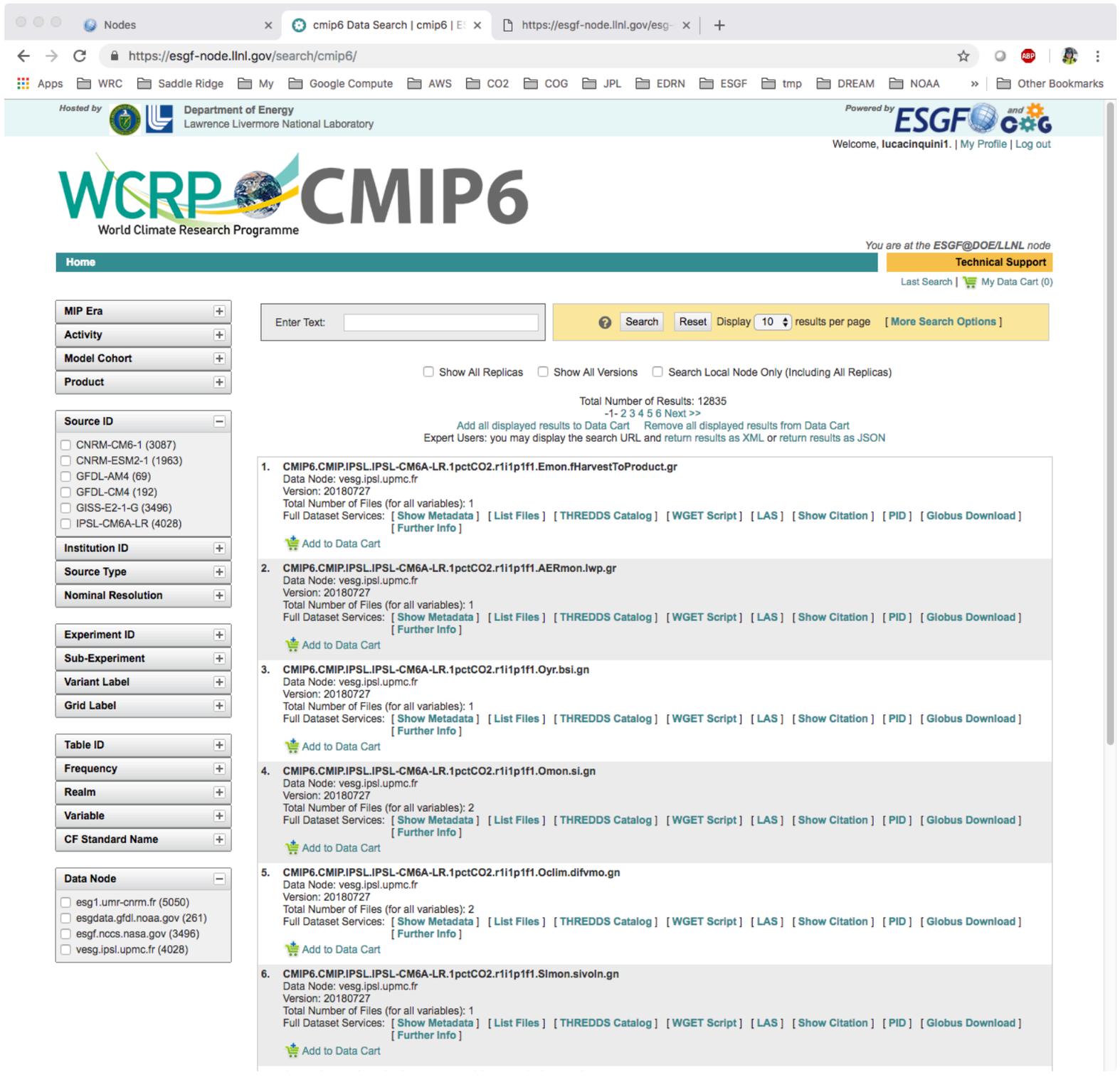
Preparations for CMIP6

- Arguably, the most important ESGF task in 2018
- Involved all ESGF developers, managers and node administrators, in particular:
  - CDNOT coordinated installation and testing of ESGF infrastructure across Nodes (S. Denvil, R. Petrie)
    - 5 “data challenges” held in 2018 to stress-test system with increasingly larger amounts of CMIP6 test data - very successful
    - Node admin guide (currently under review):
      - https://docs.google.com/document/d/1y6Nd0Bea6VC6iTguFwHZr2uMM56VSpaqaXtQxCX5t-U/edit
  - Replication Working Group is working at managing and improving the replication of core CMIP6 data across “Tier-1” Nodes (S. Kindermann, E. Dart)
    - Replica data are been published at LLNL and DKRZ with GridFTP endpoints
- WIP (WGCM Infrastructure Panel) is overseeing ESGF preparations and providing connections with the CMIP modeling groups (K. Taylor, Balaji)
  - Recently established guidelines for configuring ESGF sites hosting CMIP6 data

WCRP CMIP6
World Climate Research Programme

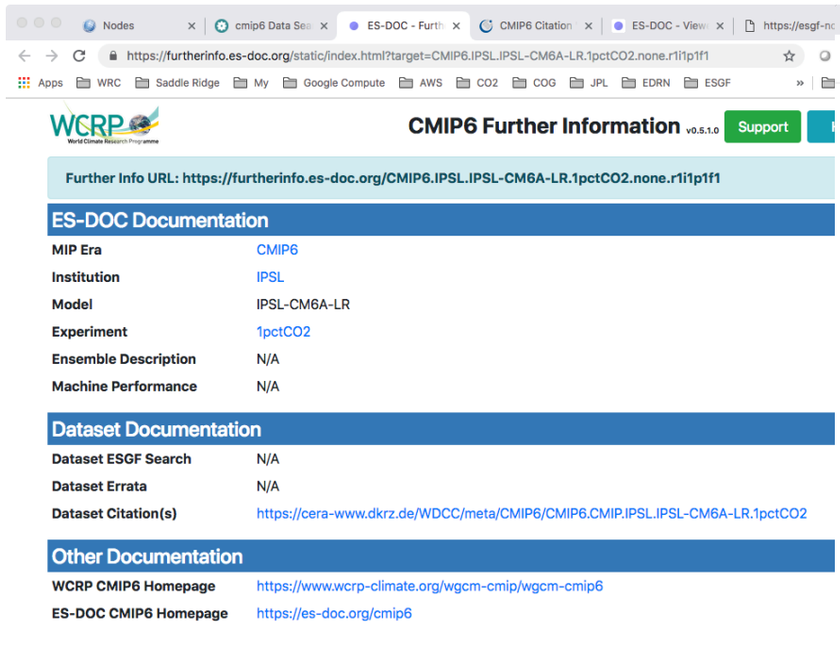
Current CMIP6 data holdings

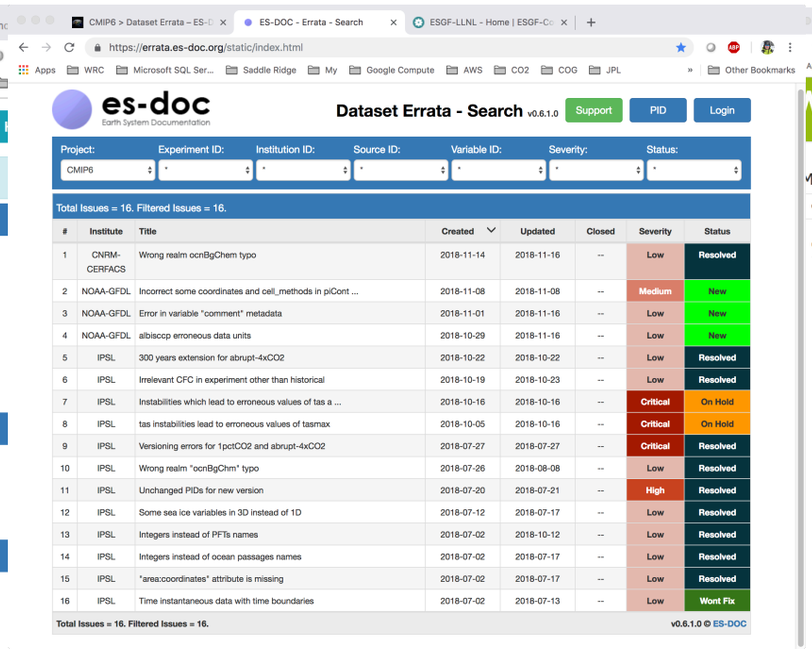
- ESGF opened for CMIP6 data in June 2018
- Currently serving CMIP6 data from 4 Data Nodes: CNRM, GFDL, NCCS, IPSL
- Data replicated at LLNL, DKRZ
- Data holdings:
  - 6 CMIP6 models
  - ~12,835 datasets
  - ~37,439 files



New ESGF Services for CMIP6

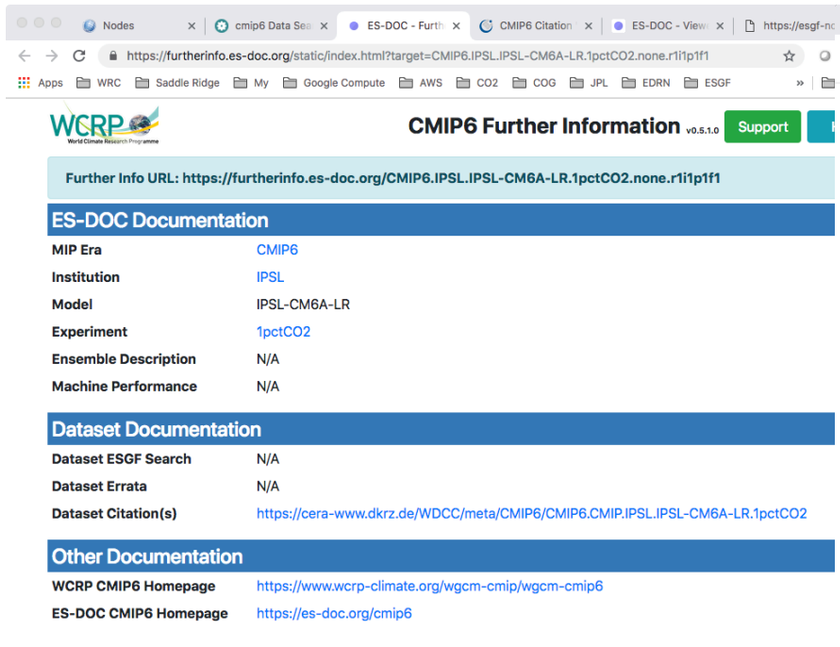
- ESGF has been working on providing new enhanced functionality in support of CMIP6
- PID ("Persistent Identifiers") service: assigns PIDs to datasets and files at time of publication for long-term identification
  - Distributed service based on RabbitMQ with 3 installations at DKRZ, IPSL and LLNL
- Suite of web hosted services that provide an unprecedented, unmatched wealth of information about data and models that generated them
  - FurtherInfo URL: embedded in NetCDF files, harvested by CoG —>
  - ES-DOC: landing pages for datasets, models, experiments, CMIP6 —>
  - Errata Service: central catalog for datasets that had to be retracted for various reasons
  - DOI Data Citation page at WDC: provides information on how to cite the data, license, content, and related datasets (forcing). Some information is encoded as schema.org —> discoverable by
  - Google Dataset Search

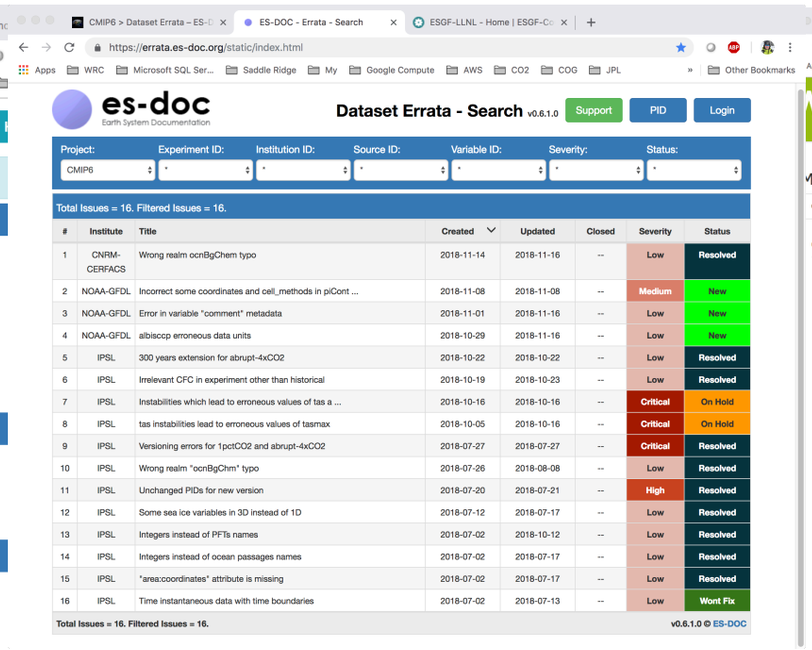




New ESGF Services for CMIP6

- ESGF has been working on providing new enhanced functionality in support of CMIP6
- PID ("Persistent Identifiers") service: assigns PIDs to datasets and files at time of publication for long-term identification
  - Distributed service based on RabbitMQ with 3 installations at DKRZ, IPSL and LLNL
- ES-DOC: eco-system of services that provide detailed documentation for models, experiments
  - Hyperlinked by FurtherInfo URL
  - Includes Errata Service - central catalog for datasets that had to be retracted for various reasons
- Citation and DOI service: provides additional information on the model, data and how to cite
- Overall, unprecedented, unmatched wealth of information about data and models that generated them

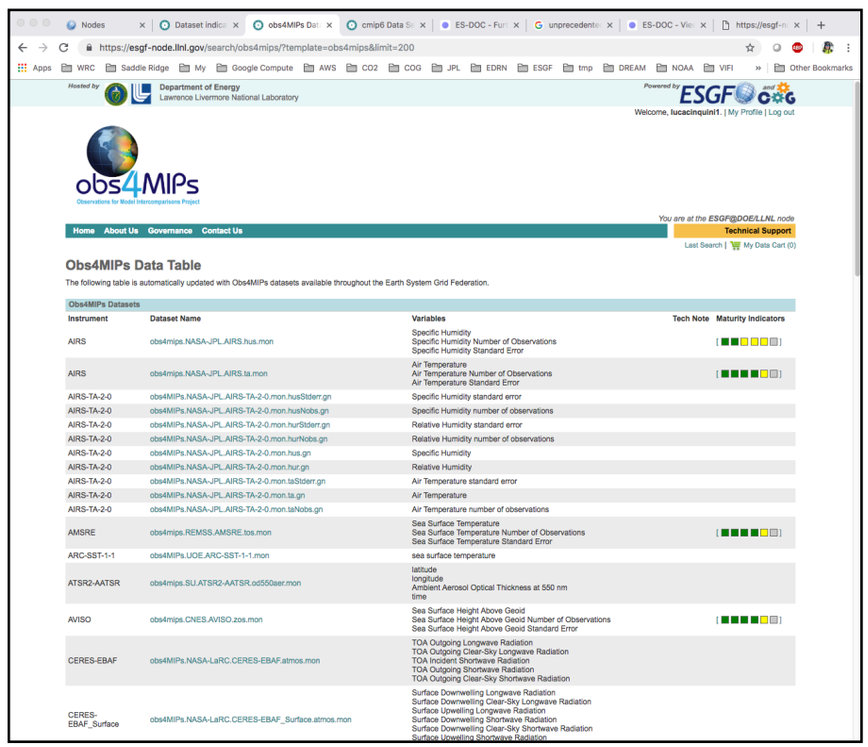




Obs4MIPs Update

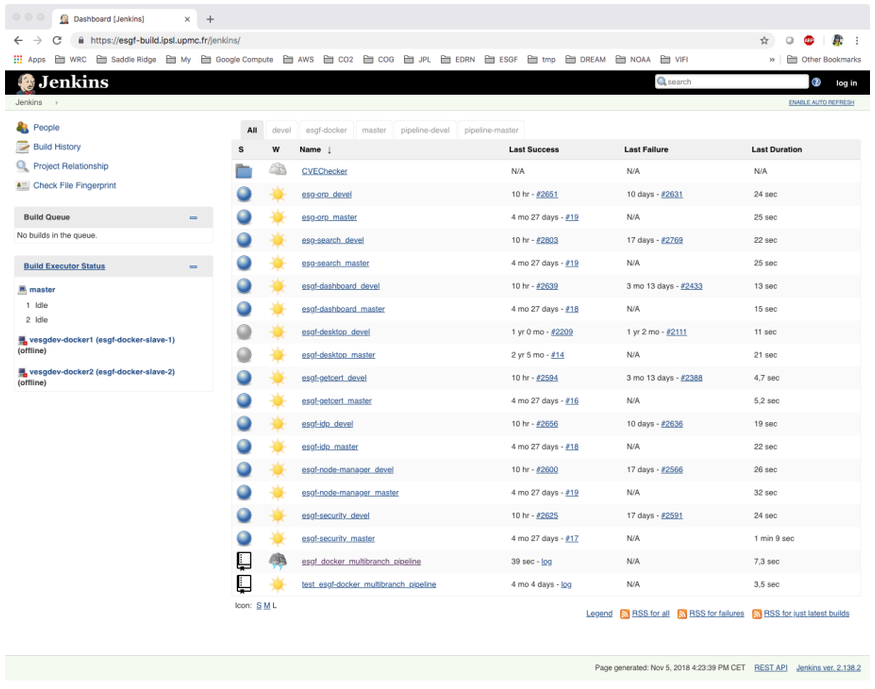
- ESGF is also increasing availability and support for observations data through its Obs4MIPs program (R. Ferraro, P. Gleckler, D. Waliser and P. Durack)
- Obs4MIPs are selected observational datasets from NASA. ESA, NOAA, etc. that follow the same data and metadata conventions as CMIP model output, intended for easier comparison and validation of model data
- Progress in 2018:
  - Additional datasets from several agencies (90 datasets total)
  - Introduced support for “dataset indicators” that represent the “maturity level” for model evaluation
  - New specification for dir structure, filenames, search facets that is aligned with CMIP6 (ODSv2.1)
- Datasets are being moved and republished at LLNL to follow the new specs
- Obs4MIPs CoG site moved to LLNL for long-term support

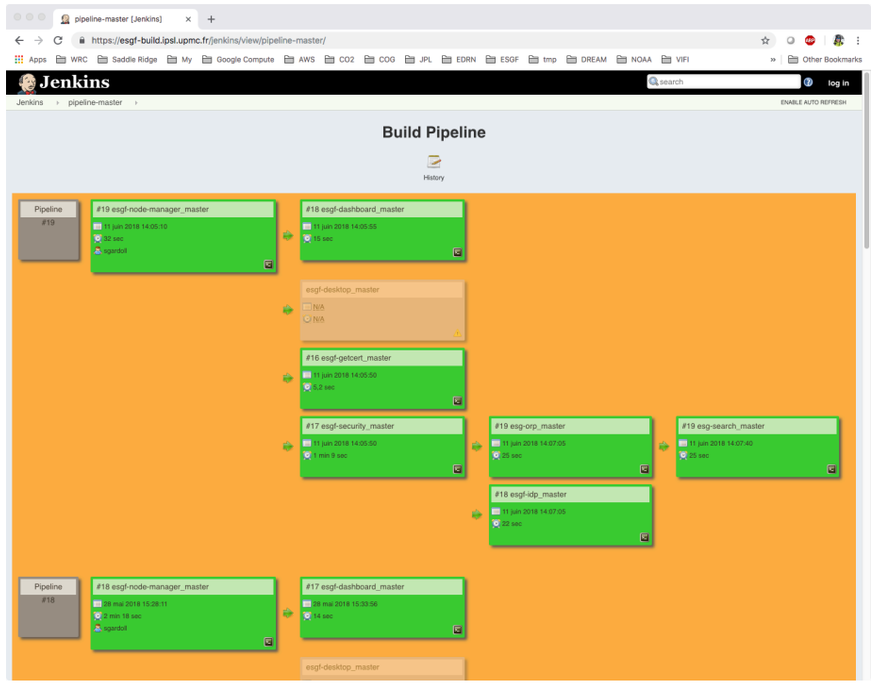
Technical Requirements		*Dataset Suitability and Maturity*			*Comparison Complexity*
Meets obs4MIPs data technical requirements	**Includes obs4MIPs technical note information**	**Closeness or robustness of measurement to observed reference quantity**	**Maturity with respect to climate model evaluation**	**Provision for robust uncertainty information**	**Complexity of Model Observation Comparison**
Data suitably processed with CMOR and/or consistent with obs4MIPs standards	Complete technical note information provided	Firmly established and/or validated methodology	Multiple peer-reviewed examples of application to CMIP climate model evaluation	Uncertainty information provided per retrieval/grid point	Comparison can be made directly with CMIP model output variable
Largely complete with minor metadata inconsistencies	Technical note information incomplete and/or could be improved	Indirect means of calculation or observations only providing partial constraint (e.g. ocean surface latent heat flux)	One peer-reviewed example of application to CMIP climate and/or examples of other sorts of model evaluation.	General uncertainty information given relative to the methodology and dataset as a whole - backed by actual field/in-situ validation exercises	Comparison requires some simple post processing of CMIP output variable(s) (e.g. vertical integral or ratio of two variables)
Non-compliant. Should be removed from database!	Technical note not provided	Largely model-derived quantity (e.g. LAI, root zone soil moisture, NPP)	As of DATE-TBS, no significant application to climate model evlauation	No uncertainty information provided	Comparison requires complex processing of CMIP output (e.g. “simulator”, budget calculation)



Installation of ESGF Node

- ESGF has worked at improving the process of installing, maintaining and securing an ESGF Node, in several directions:
  - "Classic" shell-based installer was upgraded to support 6 major ESGF releases in 2018: 2.6.5, 2.6.7, 2.6.8b, 2.6.9, 2.7.1, 2.8 (S. Ames, P. Dwarakanath)
    - New "esgf-scanner" tool that automatically creates a manifest of all software packages included in an ESGF release and lists all CVEs
    - Software stack now working on Centos7
  - Upcoming migration to new Python-based installer (3.0) (W. Hill) - now in beta
    - Will be much easier to run, test and maintain than current shell installer
- New Continuous Integration (CI) system built on Jenkins provides automatic builds and runs the test-suite for every GitHub commit (S. Gardoll)



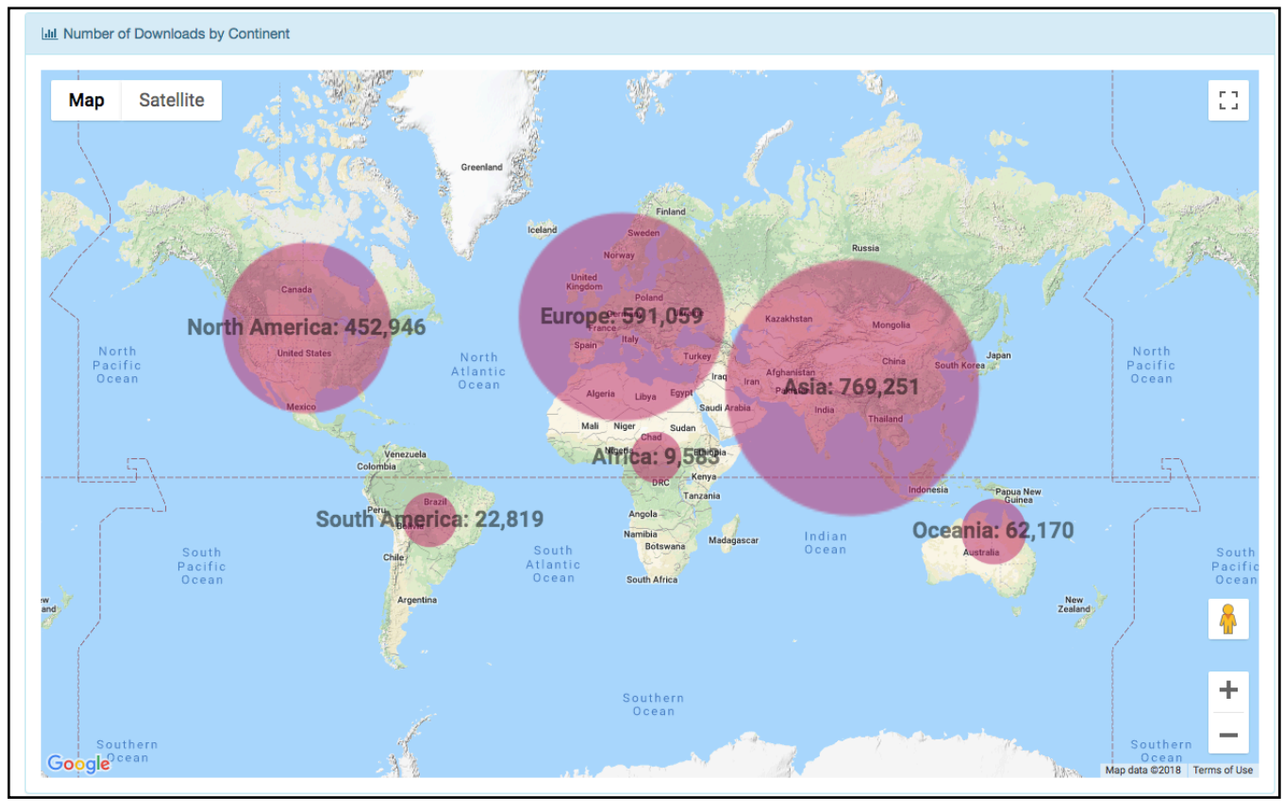


Other ESGF Development

- Idea Working Team has been progressing on transitioning the current ESGF Security infrastructure (based on OpenID 2.0) to more current industry standards: OAuth and OpenID-Connect (P. Kershaw)
  - New software components SLCS, ESGF-Auth to replace current IdP, ORP, MyProxy
- ESGF publisher and ESG prep - several upgrades to support CMIP6 and improve performance (S. Ames, G. Levavasseur)
- Dashboard team - integrating the information provider into the installer and supporting the central metrics aggregator site, also developing custom view for CMIP6 (S. Fiore, A. Nuzzo, M. Mirto)
- …and much more…

Published data over the federation

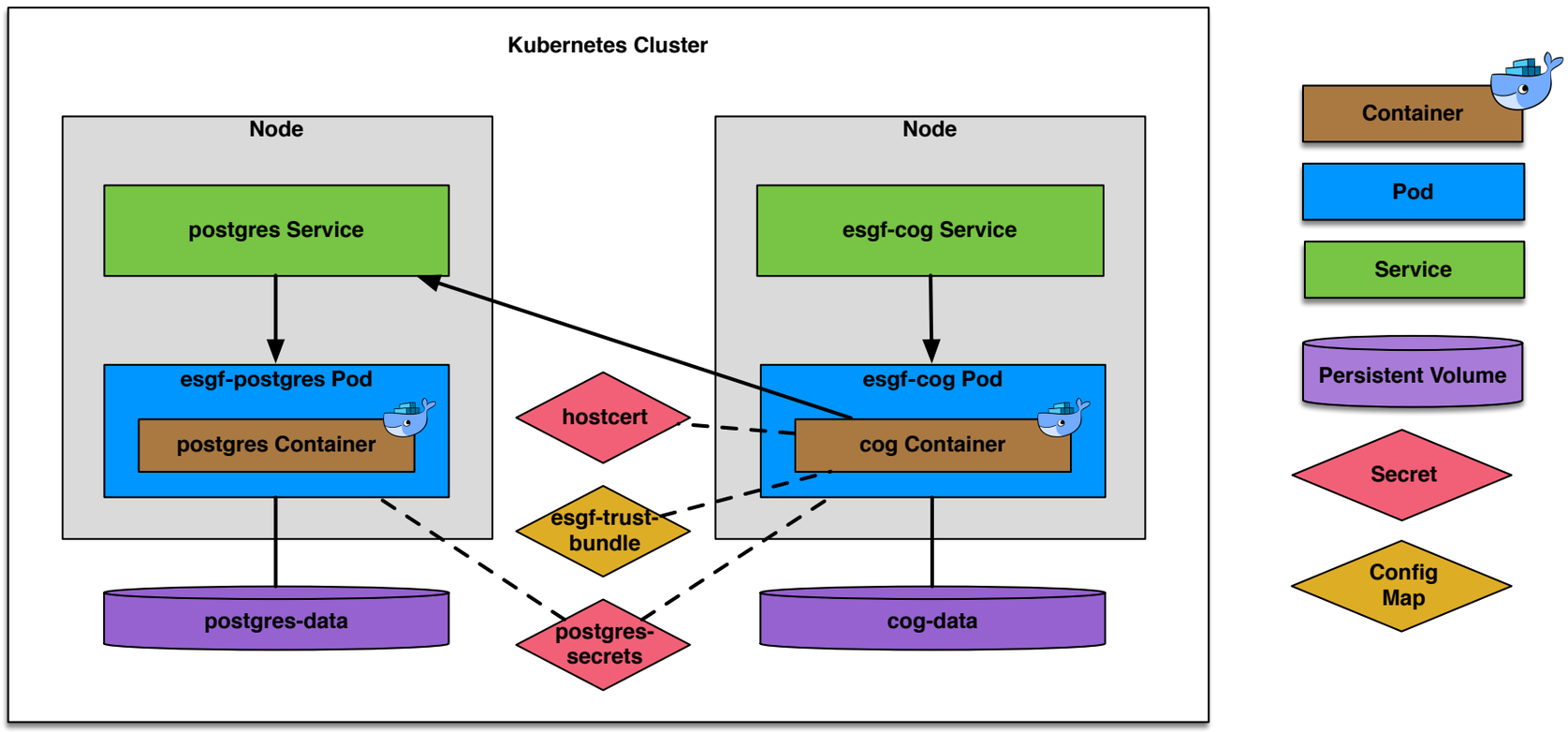
951,386 total datasets 5,823.11 TB	831,092 distinct datasets 2,725.91 TB	120,294 replica datasets 3,097.21 TB
170,605 total datasets 4,798 TB **CMIP5**	54,828 distinct datasets 1,709.53 TB **CMIP5**	115,777 replica datasets 3,088.48 TB **CMIP5**
10,821 total datasets 88.98 TB **CMIP6**	10,815 distinct datasets 88.96 TB **CMIP6**	6 replica datasets 0.03 TB **CMIP6**
		7,559 total datasets 9.28 TB **INPUT4MIPS**
3,789 distinct datasets 4.64 TB **INPUT4MIPS**	3,770 replica datasets 4.64 TB **INPUT4MIPS**	90 total datasets 0.28 TB **OBS4MIPS**



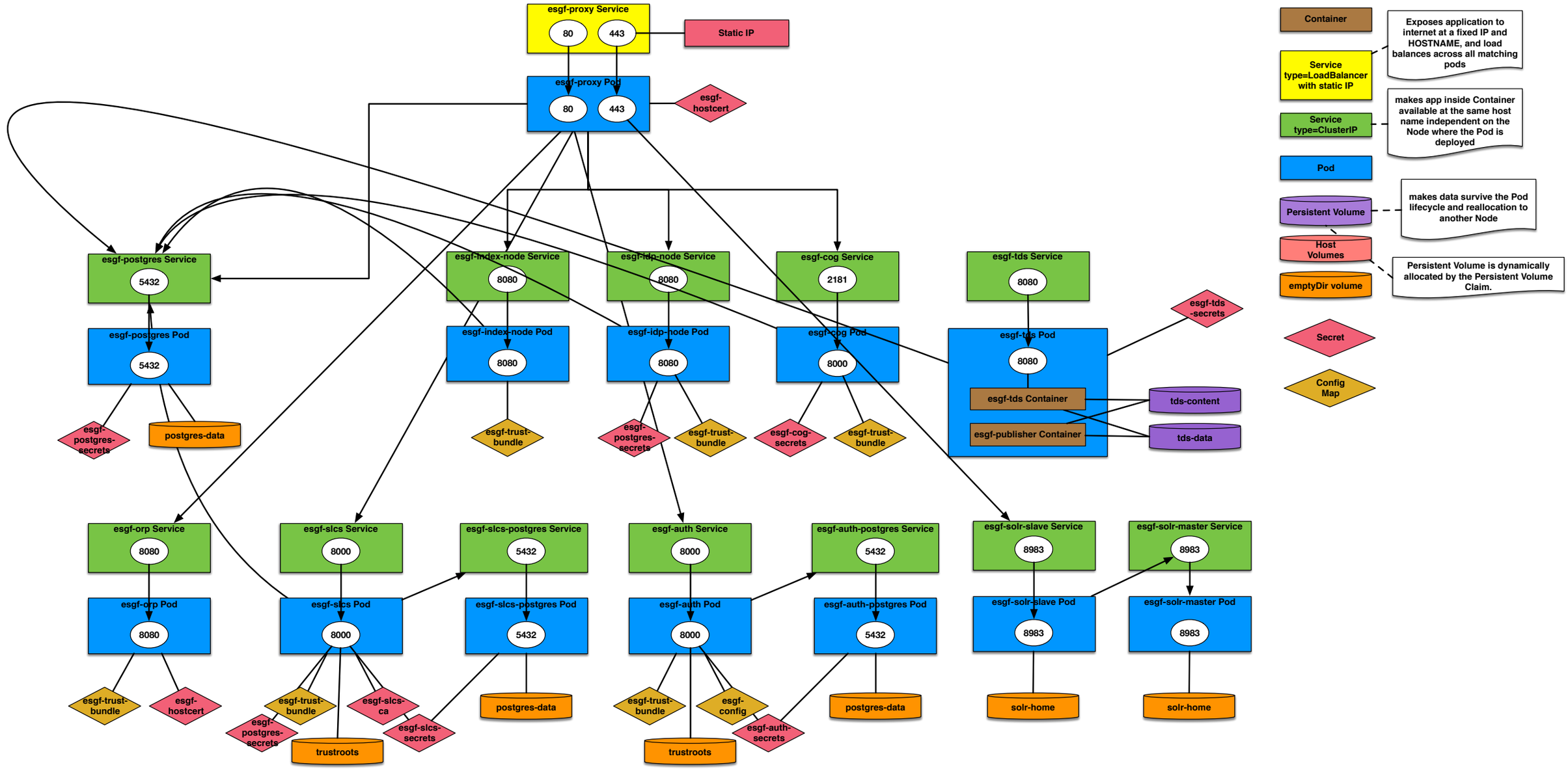
ESGF NEW DIRECTIONS

Containerization

- ESGF/Docker: alternative architecture for ESGF Node where all services are packaged, deployed and managed as Docker containers
- Advantages of container based architecture (“micro-services”):
  - Easier to deploy and test
  - More flexible
  - More scalable
  - Easier to evolve
- ESGF/Docker first release in September 2018
  - Stable but not feature complete (no Globus)
  - Based on Docker, Kubernetes and Helm

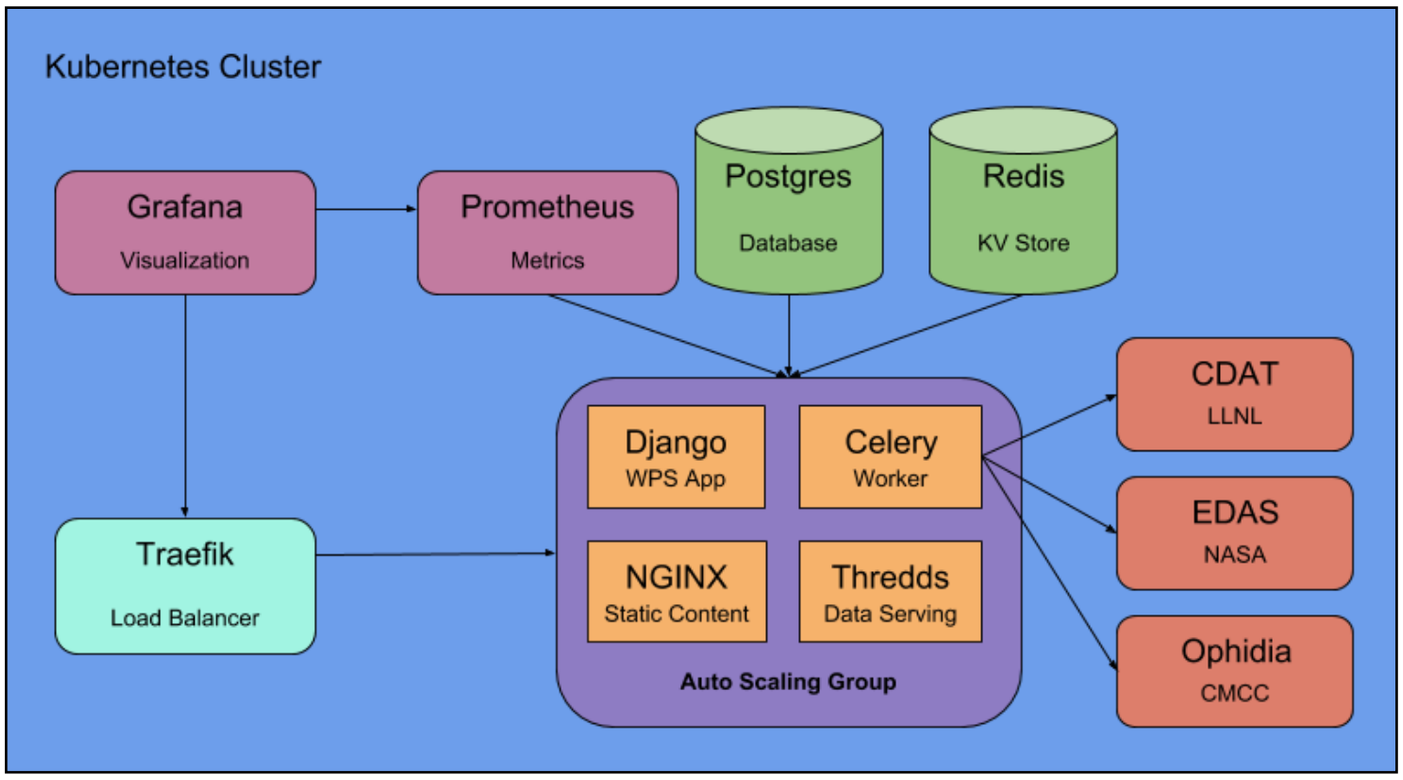


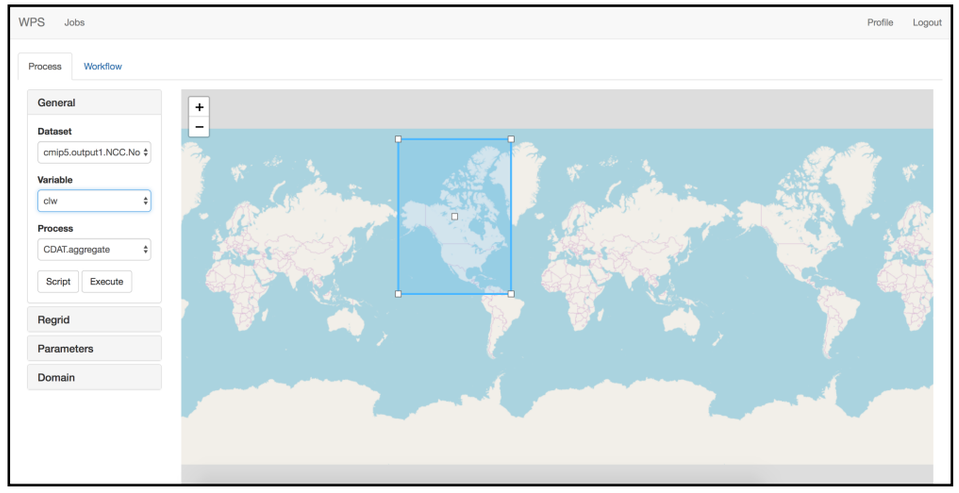
ESGF/Docker Complete Software Stack



Compute Node

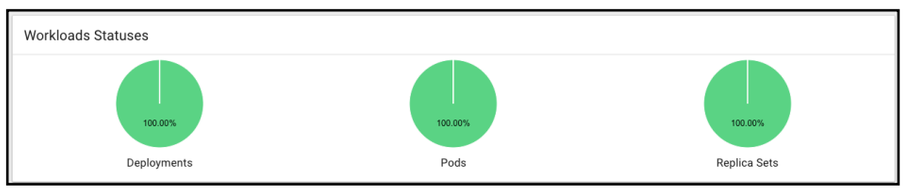
- The ESGF Compute Working Team has made great progress ind developing scalable computing capabilities for ESGF (J. Boutte, C. Doutriaux, T. Maxwell)
- Architecture of compute node was designed from the ground up as a system of interacting Docker containers
  - Highly scalable - both horizontally and vertically
- 3 alternative back-ends implement the same algorithms (sub-set, average, min/max, etc.)
  - CDAT, Ophidia, EDAS
- Status: already converted to Kubernetes+Helm, ready to be deployed alongside ESG/Docker stack

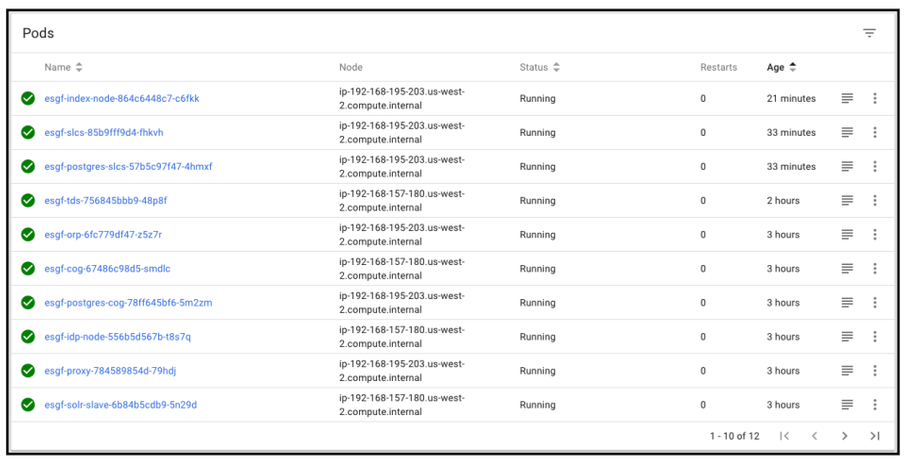




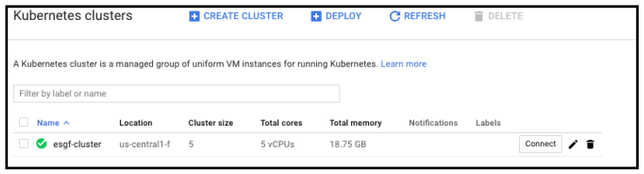
Moving to the Cloud

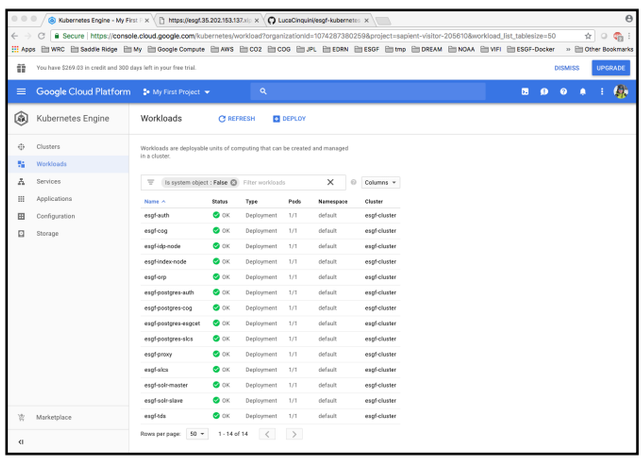
- ESGF is experimenting with moving its services -all or in part- to the Cloud
- Cloud advantages: practically unlimited scalability, high availability, managed resources
- Cloud challenges: cost model, new architecture designs
  - How to persist data, how to plan for failure, how to run on a cluster of nodes
- Several efforts undergoing:
  - ESGF/Docker with Kubernetes is immediately suitable for Cloud deployment
  - GFDL is running a prototype node on Google GKE, published some CMIP6 data, enabling access to Pangeo via openDAP
  - GSFC/JPL planning to deploy a single ESGF/NASA node on AWS GovCloud
  - New Index Node architecture based on Solr Cloud, stable deployment on ASS for several months





AWS EKS

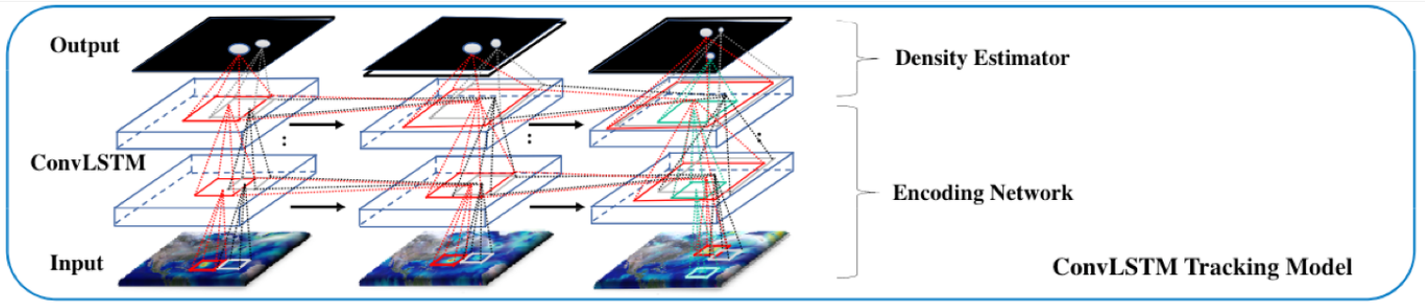


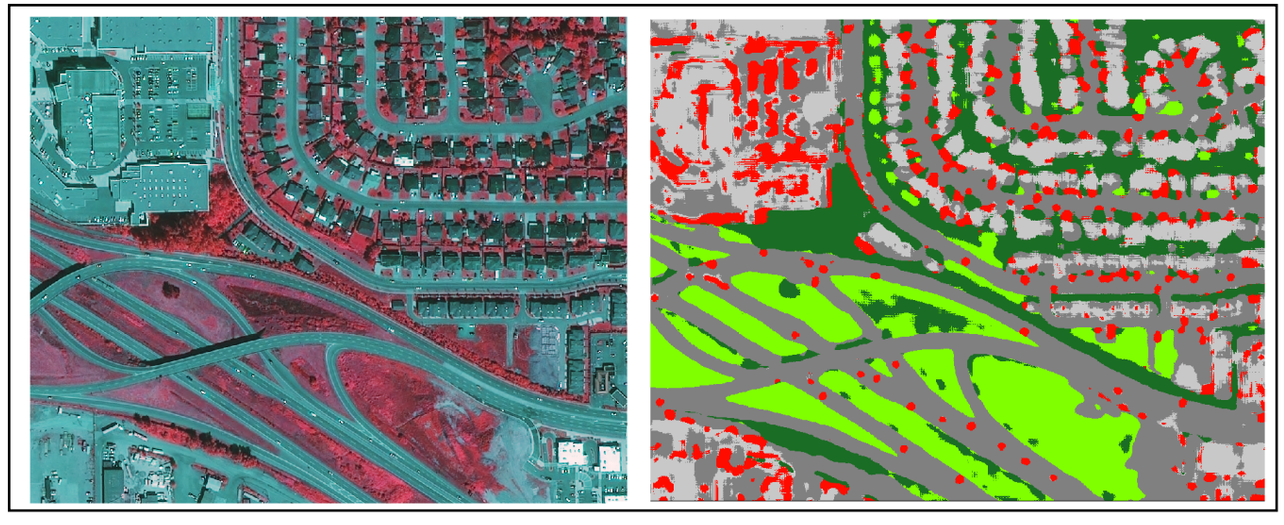


GCP GKE

Machine Learning

- How to mine the vast amounts of data held by ESGF to make reasonable predictions on future global climate and weather events?
- LLNL: “Deep Hurricane Tracker” model analyzes patterns in climate simulation data to predict hurricane tracks (S. Kim)
- CCMC: High Performance Data Analytics and Machine Learning using Ophidia - an infrastructure for executing declarative, parallel, server side analytics workflows (S. Fiore)
- CRIM: working with OGC to advance ML&DL capabilities for high resolution satellite images (T. Landry)





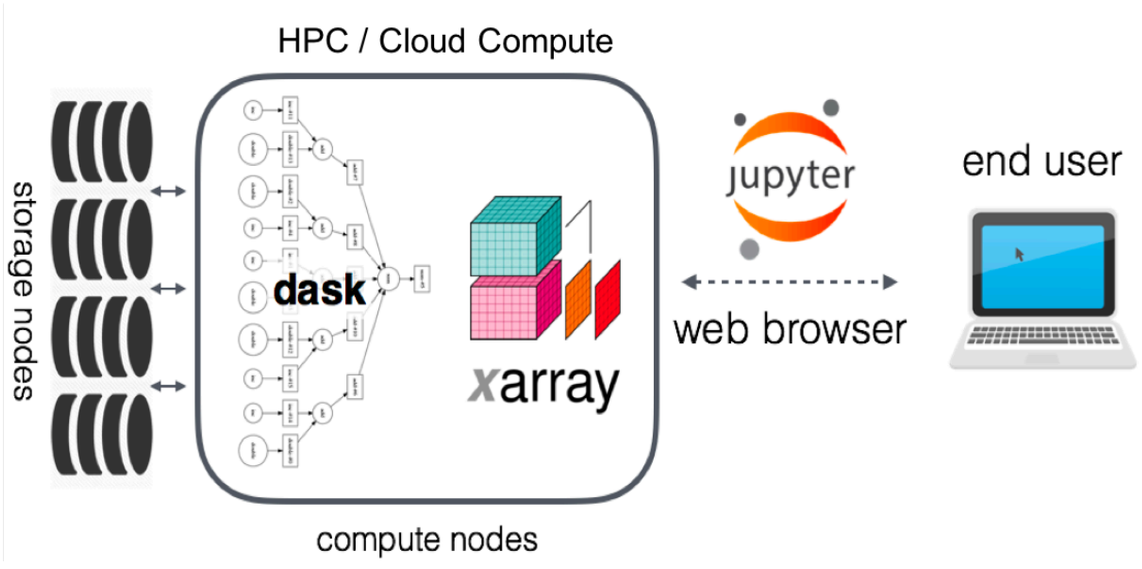
KEY CHALLENGES

Key Challenges for 2019 and Beyond

- Scalability
  - A container based architecture is highly scalable, but the application itself must be scalable
  - Must address scalability for publishing services, data catalogs (TDS), and search
- Data access
  - Making it easier for users to download data
  - Improve or replace the wget scripts
  - Better support for subsetting at the source, over space and time aggregations
- Server side distributed computing
  - Moving the computation to the data, necessitated by large Big Data volumes
  - Deploy the Compute Node operationally
  - Enable workflows that span multiple steps, at multiple sites

Key Challenges for 2019 and Beyond

- Interoperability
  - ESGF must start to interoperate with other large Earth Science infrastructures
    - NASA DAACs, also pushing towards the Cloud
    - ESA and Copernicus services (Matt Pryor)
    - Pangeo: technical and scientific community building a Python based infrastructure for scalable analysis of Earth Science data (Ryan Abernathy)
      - Jupyter for interactive analysis, Xarray/Pandas as data models, Dask for parallel computing
  - Google dataset search



Key Challenges for 2019 and Beyond

- Usability
  - Re-design the CoG UI to remove some un-needed functionality
  - Mention the main topics coming up in email support requests
  - Possibly establishing and supporting a help desk

Conference Highlights

- Day 1:
  - Keynotes and general directions from funding agencies and major stakeholders
  - Computing services and Machine Learning
- Day 2:
  - ESGF support for CMIP6
  - Interoperability with other Earth System infrastructures
- Day 3:
  - Reports from ESGF working groups
  - Open forum on re-thinking the ESGF architecture
  - Awards Ceremony
- Day 4:
  - ESGF-XC meeting & working groups parallel sessions
  - Conference summary
- Poster session and live demos at the end of each day

Earth System Grid Federation